# EXHIBIT F

(Part 2)

Page 111

events, recorded some information which would lead him to conclude that the robots.txt exclusion pages between 4500 and 5,000 bytes? Do you know that?

A. I don't know whether that's the case or not, but that's not what I was saying before. I was talking about whether there were other files of the same size.

Q. Yes. That's not my question. My question is a little bit more precise. I am asking you to consider whether the bases for Mr. Lenky's analysis -- and I'll go through them in more exhaustive detail if you like -- number one, that robots.txt exclusion pages are between 4500 and 5,000 bytes, number one, and that by filtering out the HTTP access logs for successful data transmissions between Internet Archive and the defendants where the transmission size was in range, that same range, and then exclude image files and the robots.txt file, a subset of log file entries representing potential robots.txt exclusion page views, etc. -- whether, given that this information was based upon Mr. Mohr and his knowledge, whether that would change your concern about the conclusion that was reached.

Page 112

MR. LEWIS: Objection to the form.

A. No. What I'm saying is that the numbers that Mr. Lenky concludes here do not follow from the assumptions and assert -- the facts he claims here in this paragraph. There are other things one would have to assume to reach -- to reach the numbers that he reaches, and he does not state those other assumptions, let alone giving any reason for why they would be true.

Q. Well, what, in your view, are the fallacies with his reasoning in this area?

A. The fallacy is assuming -- one fallacy is assuming that there could not have been any file between 4,500 and 5,000 bytes in size other than the ones that he specifically excluded.

Q. What do you mean more precisely than that?

A. Well, he gets to his final numbers by subtraction. He first identifies all of the accesses that were in that size range, and then he subtracts off some files that were of that size range that are not -- that are not exclusion pages. In order to get to the accurate number, he would have to exclude all of the pages of that size that were not exclusion files. He subtracts

Page 113

off some of them, but he --

Q. But you're not -- I don't --

A. -- but he doesn't say that he's subtracting off all of them. He doesn't give any reason for believing that he subtracted off all of them. There are other categories of pages that he has not excluded here that would be included in this final number.

Q. Anything else?

A. Well, that's certainly the primary -- the primary reason to question his numbers here.

Q. Well, if there are any secondary reasons, I want to know about them. Is there anything else?

A. Nothing else comes to mind.

Q. Looking at the next paragraph -- if you want to take a moment to read it, please do. Have you read through that, the next paragraph?

A. Let me take a minute.

Okay.

Q. Do you have any dispute as to the methodology that he used in that paragraph to confirm the results and conclusions from the prior paragraph?

A. I understand the methodology he used.

Page 132

Q. Did you otherwise, through any other source, come to a conclusion as to which individuals may have been directly involved?

A. No.

Q. I think you testified earlier that you did not speak to anyone at Harding Earley at any time for any purpose. Correct?

A. That's correct.

Q. Do you disagree with Mr. Lenky's conclusion that the two distinct user-agent strings present in the web logs indicate that there were at least two separate machines who were responsible for initiating their requests for the content?

A. I agree that different user-agent strings would tend to indicate different computers. As to whether there were two different strings, I would have to look to the logs.

Q. Look at the logs.

A. In a quick scan of the logs, I'm only seeing one, but perhaps there's information you can point me to.

Q. I'll be happy to. Can I have that exhibit, please?

A. Sure.

Page 133

Q. Referring you to the first page of the exhibit, the first log entry and the third full log entry from the bottom.

A. Yes, I see those.

Q. Those would be two distinct user-agent strings?

A. Yes, those are different user-agent strings.

Q. That would seem to suggest two different machines were involved?

A. Yes, that would tend to indicate that.

Q. Look at page 43 of your report.

A. You mean paragraph 43?

Q. Yes, paragraph 43. Sorry. You mention in this paragraph that they, the Harding Earley employees, did not engage in "hacking." How do you define "hacking" in this context?

A. Well, I'm using the term here to mean activity that is -- activity that is devious and out of the ordinary, essentially.

Q. In your mind, hacking connotes an intent to be devious?

A. Well, the term "hacking" is used in many different ways within computer science. It's sometimes used to refer only to someone exhibiting

Page 140

following testimony:

"QUESTION: Is it inconsistent with your review of the logs that the following happened, number one, that members of the Harding Earley law firm repeatedly attempted to access an archived web page of Healthcare Advocates within a short span of time?")

A. They did request archived pages from the Internet Archive. And depending on how short a span of time and how many repetitions you're talking about, I suppose you might be able to say that they requested -- they made repeated requests over a period of time.

Q. Over a short period of time?

MR. LEWIS: Objection.

Q. When I say a "short period of time," I mean a matter of minutes. Did you observe that through the logs?

A. There are points in the log where one sees more than one request in a period of a few minutes.

Q. For the same archived content. Correct?

A. For the same archived content, I'd need to review the logs to answer that.

Q. Please do.

Page 141

Have you had a chance to review the logs, Professor?

A. Yes.

Q. Can you answer the question now?

A. I do see some instances of the same content being requested close together in time.

Q. Within a matter of minutes. Correct?

A. Yes.

MR. LEWIS: Objection to form.

A. Yes.

Q. And do you also note that there's a pattern throughout the logs whereby there are a number of unsuccessful requests for particular content followed by a successful request?

MR. LEWIS: Objection to form.

A. I don't know whether that's in the logs or not, and I don't know that I can answer the question at all by looking at printed-out logs.

Q. What else would you need to review in order to rely on that?

A. I think I would need to use pattern-matching tools and so on, to have an electronic version of the logs and to operate on that.

Q. Did you use your pattern-matching tools

Page 143

examples. Generally trying to classify the requests and to count different -- the different kinds of events.

Q. During the break when you were reviewing the logs in order to answer the question that I initially asked you, I believe you said that you observed that there were instances within a matter of minutes where there were repeated requests for the same archived web content. I don't want to misstate your testimony, but is that accurate?

A. More than one request for the same content during a matter of a few minutes, yes.

Q. Did you also observe that at the end of that state of requests, there was a successful access of that same requested web content?

A. No.

Q. What did you observe?

A. I didn't try to look for the end of the chain or the sequence.

Q. Could you do that if you sat here today?

A. It would take some time. The logs in this form are not easy to digest.

Q. Do you recall at any point in time in reviewing the logs seeing at least once that same pattern, repeated attempts to access a specific

Page 144

archived web page which were unsuccessful, followed by a success in achieving access to that web page within a short period of time, meaning a span of minutes?

A. No, I don't recall seeing that, that pattern.

Q. Did you specifically look for that pattern?

A. I don't recall looking for it.

Q. Let me show you again what's been marked Bonini-8, the robots exclusion page. At the same time I'll refer you to your report, paragraph 49 and 50.

Starting at paragraph 49, it's your opinion that the law firm's accesses to the Wayback Machine were not unauthorized, and, moreover, that there was no intentional unauthorized access on their part?

A. Well, I disagree with Mr. Lenky's apparent assertion that the law firm employees knew that what they were doing was unauthorized.

Q. Okay. Well, then, I understand what you're saying. I think that's different from what I'm saying. I'm asking you two discreet questions, and I'll ask them to you separately in

Page 146

Q. I'll rephrase the question if you can't answer it.

A. Well, it does appear that Healthcare Advocates during at least part of the relevant time had a robots.txt file in place that asked the Internet Archive crawler not to visit their page. As to what their intention was, I don't know.

Q. Is it also clear to you that the robots exclusion that Healthcare Advocates had in place was effective in blocking third-party access to the archived web content during that period of time?

MR. LEWIS: Objection to form.

A. Was it effective? Well, we know that some accesses to the archived content were successful.

Q. But don't we also know that --

MR. LEWIS: I don't think he's finished his answer. Have you finished your answer?

MR. CHRISTIE: All right. I'll let him finish.

A. I'll stop there.

Q. Well, don't we also know that some attempted accesses were rebuffed?

A. Some attempted accesses were

Page 147

unsuccessful, yes.

Q. And that was based upon the robots exclusion. Correct?

MR. LEWIS: Objection to the form.

A. Well, that was the behavior of the Wayback Machine when those requests were made.

Q. So would you agree with me that at least in part of the time during the period July 9, 2003, through July 14, 2003, it was a properly configured and properly implemented robots.txt exclusion on the Healthcare Advocates web server?

MR. LEWIS: Objection to form.

A. This was probably the case for at least part of that period.

Q. And is it possible that it could have been that case for the whole period of time?

A. That is possible.

Q. Is it also accurate to say that, based on your report, you dispute Mr. Lenky's conclusion that any access -- that any unauthorized access, to the extent that there is any, by the law firm representatives was intentional?

MR. LEWIS: Objection to the form.

A. Well, Mr. Lenky doesn't explain his reasoning, if I recall correctly, for why he says

Page 149

file itself and the query exclusion page, which we marked as Bonini-8. Is that accurate?

A. Yes, that's roughly what my report says.

Q. What you don't talk about in your report, number one, is the number of instances that the Harding Earley representatives actually physically and intentionally sought out and viewed the robots.txt file, do you?

MR. LEWIS: Objection to form.

A. Well, I do say that he makes -- one of the bases he appears to rely on is Harding Earley employees seeing the robots.txt file. And if he -- and to the extent that he relies on how many times that happened, that's something that I'm discussing here.

Q. Well, please look at page 7 of his report. There's a table that goes through to page 9 which sets forth specific instances where the Harding Earley representatives are requesting the robots.txt file. Do you see that table?

A. Yes. He points to one user doing it once and another user, he says, doing it seven times.

Q. Did you have occasion to, in the course of your analysis, to compare the claims in this

Page 150

table with the logs themselves?

A. I don't recall whether I did or not.

Q. Do you have any reason to dispute the accuracy of the information in this table?

A. Sitting here today, no.

Q. Do you see that the -- that in the table in Mr. Lenky's report, that the timing of the requests for the robots.txt file are early in the day on July 9? When I say "early in the day," I mean between 6:25 Pacific time and 8:09 Pacific time. Do you see that?

A. Well, I see the times. I'm not positive about the time zone. It says "minus 0700" on the time zone.

Q. Well, I thought you didn't dispute the accuracy of the table, which includes --

A. Well, there are times listed here. It says, for example, on the first access the time as 06:25:20. Then it says "minus 0700," which I know denotes a time zone, but I'm not sure which time zone it denotes.

Q. Okay. Do this, please. Take the logs, ones for Internet Archive, which are marked Mohr-2, and please look and see if you can find the first referenced access in the logs at 6:25,

Page 151

6:25:20.

A. Yes, I see it.

Q. Do you want to take a look to see if you can find the other ones, too?

A. Sure. Are you asking me to confirm the things on the right-hand side of the table or that there accesses to robots.txt at the time listed on the left?

Q. While you're at it, I want you to confirm everything.

A. Okay, I've reviewed all of these.

Q. Having reviewed in Mohr-2 the items in the table on pages 7 through 9 of Mr. Lenky's report, do you dispute any of the information contained in the table as per your comparison with the logs?

A. There is, I think, one error in the table --

Q. What's that?

A. -- which is in the second access for user B.

Q. Yes.

A. I believe he has the time slightly wrong. It's off by a few seconds. Also, I should note that the language above the table that says

Page 152

that "Each of these is an access to the robots.txt file" is not correct. The last entry in the table is not a reference to the robots.txt file. The last entry in the table is not a request for the robots.txt file. It's a request for other content.

Q. Isn't that what the explanatory note --

A. Yes, his explanatory note says that, but the caption above the table says that -- this says that the table is a list of requests for robots.txt file, so that doesn't match. Otherwise, these appear to be as described in the table.

Q. You agree with his interpretation of the significance of the log entries in the right-hand column. Correct?

MR. LEWIS: Objection to the form.

A. His descriptions in the right-hand column appear to be accurate. As to what interpretation he makes of that, that's perhaps a different story. Focusing on just he wrote in the right-hand column, yes, that does appear to be consistent with the logs.

Q. So you would agree, then, that early in the morning on July 9 of 2003 there were repeated

Page 153

requests specifically for the robots.txt file emanating from the Harding Earley law firm?

MR. LEWIS: Objection to form.

A. There were requests for the robots.txt file from the Internet Archive.

Q. And, in fact, there were requests not only for the current version of the robots.txt file. Correct?

MR. LEWIS: Same objection.

A. Correct.

Q. There were requests for versions from the year 2000. Correct?

A. Yes, there appears to be the requests for the version from the year 2000.

Q. As well as a request for all archived versions of the robots file. Correct?

MR. LEWIS: Same objection.

A. Yes, there appears to be such a request.

Q. So that would appear to indicate that there was at least one user agent who was conducting an investigation into the robots.txt file?

MR. LEWIS: Objection to the form.

A. There was this one user agent who -- which might or might not have been a single person

Page 154

or more than one person. It appears to be one computer from which these requests were made.

Q. Does it also appear that the requests were made for the robots text file from the query exclusion page, Bonini-8?

MR. LEWIS: Objection to form.

A. That I can't be certain of.

Q. Why can't you be certain of that?

A. Well, the question is whether the previous page that the user saw was an exclusion page or not, and that is not immediately evident from the log.

Q. Okay. Well, if you look at the text string in the bottom left-hand corner of Bonini-8 and you compare it to the right-hand column of the table in Mr. Lenky's report, would that alter the answer you just gave me?

A. No.

Q. Why not?

A. Because the requests that are made for archived content would yield either an exclusion page for either content or a list of content that's available. And you cannot immediately tell from looking at the URL, which was provided in a given instance. So I can't tell, sitting here,

IN THE UNITED STATES DISTRICT COURT
FOR THE EASTERN DISTRICT OF PENNSYLVANIA

HEALTHCARE ADVOCATES, INC.,

Plaintiff,

v. Civil Action No. 05-03524

HARDING, EARLEY, FOLLMER & FRAILEY,

Defendant.

## EXPERT REPORT OF EDWARD W. FELTEN

### I. Introduction

1. My name is Edward W. Felten. I am a Professor of Computer Science and Public Affairs at Princeton University, where I have taught for thirteen years. I also serve as the Director of Princeton's Center for Information Technology Policy.

2. My research and teaching in computer science have focused on computer security and privacy, Internet software, computer systems, and technology law and policy. I have won awards for my work, including a National Young Investigator Award from the National Science Foundation and an Alfred P. Sloan Fellowship. In 2004, Scientific American magazine named me to the Scientific American Fifty, its

EXHIBIT
Felten-5
2-2-07

list of fifty global leaders in science and technology. A copy of my Curriculum Vitae is attached to this report as Appendix A.

3. I am being compensated at a rate of $700 per hour for my time working on this matter, and my out-of-pocket expenses are being reimbursed.

4. I have served as an advisor on information technology issues to many agencies of the U.S. government, including the departments of Justice, Defense, and Homeland Security; and the Federal Trade Commission. I have testified before committee hearings of the U.S. Senate and U.S. House of Representatives.

5. I was the lead technical expert witness for the Department of Justice in the antitrust case *U.S. v. Microsoft*. In that capacity I testified twice, in December 1998 and June 1999, before Judge Thomas Penfield Jackson in the District of Columbia. I testified briefly in 2000 in *Universal v. Reimerdes* before Judge Lewis Kaplan in the Southern District of New York. I testified twice, in 2000 and 2003, in the patent case *Eolas v. Microsoft*, before Judge James Zagel in the Northern District of Illinois, Eastern Division. I testified in 2006 in *ACLU v. Gonzales*, before Senior Judge Lowell Reed in the Eastern District of Pennsylvania.

6. I have written and taught extensively about Internet software and Internet protocols. For example, I have taught Princeton's course on computer networks, which covers many Internet protocols.

7. My research and teaching on computer security has included issues of filtering, access control and circumvention. For example, I have taught about these topics in courses on information security and information technology policy.

8. I have been asked to render my expert opinion on certain technical issues relating to the claims in this matter, and to respond to the expert report of Gideon Lenkey. My opinions are based on my training and experience, and on materials I have consulted. I have consulted the following materials: all materials referenced in this report, and Web server log files from HCA and IA.

## II. Internet and Web Technology

9. The Internet is a global network of computers constructed by patching together many local area networks that use widely varying communication media such as telephone lines, dedicated data cables, and wireless links. The Internet is characterized by its global scope and by the use of certain standard data formats and protocols such as the Transmission Control Protocol and the Internet Protocol (together, "TCP/IP") that ensure that any two computers on the Internet can exchange information with each other.

10. Computers on the Internet are named or addressed using two types of designators: IP addresses and DNS names. IP[1] addresses are used by the Internet's infrastructure to route traffic to a computer, just as telephone numbers are used to route calls to a phone. IP addresses are written in numeric form, such as 207.41.14.192. DNS[2] names utilize a more user-friendly textual format; an example is www.uscourts.gov. DNS names are translated automatically into IP addresses, so that users can take advantage of the friendlier DNS naming system while the

[1] IP stands for "Internet Protocol," which is the basic mechanism used to enable communication on the Internet.

[2] DNS stands for "Domain Name (or Naming) System," a system for giving computers on the Internet user-friendly textual names and translating those names into corresponding IP addresses.

Internet's basic infrastructure relies only on the more machine-friendly numeric IP addresses.

11. The Internet serves as the basic communication infrastructure for a wide range of electronic information services and activities, including electronic mail, electronic discussion groups, teleconferencing, remote access by traveling workers to institutional databases, and the World Wide Web ("the Web"). Although the Web uses the Internet as its basic communication infrastructure, the terms "Internet" and "Web" are not entirely synonymous.

12. The Web is characterized by a set of standard data formats, including HyperText Markup Language ("HTML"), and a set of standard communication protocols, such as HyperText Transfer Protocol ("HTTP"), that together allow computers to publish and view Web "pages" that may contain links to other such pages. The Web is made up of the global collection of pages that meet these format specifications, along with the global collection of computers that store and transmit pages on demand via these standard protocols.

13. A Web browser is a computer program that allows a user, typically sitting at a personal computer or laptop, to access and read Web pages. Popular Web browsers include Microsoft's Internet Explorer, and the Mozilla project's Firefox.

14. A Web server is a computer that stores Web pages and makes them available across the Internet using Web protocols such as HTTP. The term "Web server" is also used sometimes to refer to a software program that makes this possible.

15. Web pages are identified by Web addresses such as http://www.paed.uscourts.gov/contents.asp (the address of the table-of-contents page for the Federal Courts in the Eastern District of Pennsylvania). These addresses are technically known as Uniform Resource Locators ("URLs"). A URL designates a Web server ("www.paed.uscourts.gov" in this case), a particular page on that Web server ("contents.asp" in this case), and a protocol to use in retrieving the page ("http" in this case).

16. Web servers normally keep log files, which record every request made to the server. A Web server log consists of a sequence of entries. Each entry typically records information about one request, such as the date and time, the specific page or file requested, the address of the requesting computer, the outcome of the request (success, or the type of error that occurred), and the amount of data transferred in satisfying the request.

### A. Caching

17. When a computer program accesses a Web page, the computer will sometimes store a copy of that Web page, in case the page is needed again later. This is known as "caching." If the program needs the page again, it can get it from its local cache, rather than having to contact the Web server again to get another copy of the page.

18. Caching has advantages – for example, retrieving a page from the local cache is faster than downloading it across the Internet – but it has drawbacks too. One drawback is that if the page changes on the original server, the cached copy will be out of date, and a program that relies on the cached copy will have outdated

information. For this reason, cached files are normally timestamped when they are initially put in the cache, and are discarded after some time interval has passed. For example, if cached files are discarded after twenty-four hours, then information retrieved from the cache will never be more than twenty-four hours out of date.

### B. Web Crawlers

19. A Web crawler is a computer program that catalogs Web pages. Crawlers try to discover as many pages as they can, and they download and store copies of the pages they discover. In other words, a crawler makes an archival copy of whatever portion of the Web it can discover.

20. Crawlers are used by search engines, such as Google and Yahoo, to help people find information on the Web. For example, if you type "Edward Felten" into Google, its response will point you to one of my Web pages. It can do this because it has archived the contents of that page (and many millions of other pages) and built an index of its archive, thereby allowing it quickly to find pages containing particular words.

21. Crawlers typically operate on their own, without a human sitting at the computer. The basic mechanisms of the Web, such as the HTTP protocol that is used for transferring most Web pages, are designed for computer-to-computer interaction. When a user is browsing the Web, his Web browser – a computer software program – is accessing Web pages on his behalf. In the same way, a program like a crawler can access Web pages, even if no human is present. (A human would have programmed the crawler (i.e., given it instructions) beforehand.)

22. Crawlers are a well-known and accepted part of the Web "ecosystem." Many different crawlers exist.

### C. Robots.txt Files

23. For various reasons, Web site publishers sometimes prefer that their sites, or certain parts of their sites, not be visited by crawlers. Many of the people running crawlers are happy to comply with publishers' requests not to crawl their pages.

24. In about 1993, technologists recognized that it would be useful to have a standardized method by which site publishers could ask crawlers not to visit their sites. Ensuing discussions led to a consensus result known as the "robots.txt standard." It is summarized in an online document entitled "A Standard for Robot Exclusion", available online at http://www.robotstxt.org/wc/norobots.html (hereinafter, "Standard for Robot Exclusion").

25. This "standard" was not drafted or ratified by any formal standards body but was "ratified" only by common usage. The Standard for Robot Exclusion summarizes its status as follows:

> It is not an official standard backed by a standards body, or owned by any commercial organisation. It is not enforced by anybody, and there no guarantee that all current and future robots will use it. Consider it a common facility the majority of robot authors offer the WWW community to protect WWW servers against unwanted accesses by their robots.

(Standard for Robot Exclusion)

26. If a Web site publisher wishes to use the robots.txt mechanism, he creates a file called "robots.txt" and has his Web server offer this file for download. This file, if present, contains a series of records. Each record gives the names of one or

more crawlers, and enumerates the parts of the web site that the specified crawlers are requested not to visit.

27. The robots.txt file can be viewed by people, but it is intended to be read by crawlers. A crawler can look for the robots.txt file and (assuming it is present).can download and interpret it. The crawler can look for records that name it, and by reading those records it can learn that the web site author is requesting that it not visit certain parts of the site.

28. If no robots.txt file is present, it is assumed that the Web site publisher welcomes crawlers.

29. Many crawlers are programmed to read the robots.txt file and comply with the requests therein. Doing so is considered polite behavior.

30. The state of a site's robots.txt file (i.e., whether the file exists, and if so what it contains) may change from time to time. The state at some point in time communicates the site publisher's requests as of that time. Requests are not retroactive. If I add a request to my robots.txt file at noon today, that request applies to crawler visits from noon today onward. It does not apply retroactively to requests made before noon today. It does not mean that crawlers' requests made before noon today were unwelcome, nor does it mean that crawlers should not store or use information gathered from my site before noon today.

31. If a site publisher changes his mind about whether he wants a crawler to access his site, he can modify his robots.txt file accordingly. For example, if he is currently asking a crawler not to visit the site but he wants to welcome that crawler in the future, he can remove or modify the lines in the file regarding that crawler. A

crawler may check the robots.txt file from time to time to see whether the site publisher has changed his mind.

32. A web site author can change his robots.txt file at any time, and the new version of the file is assumed to apply as soon as it is made available. The contents of my robots.txt file as of noon today do not specify my wishes as of tomorrow morning – to know my wishes tomorrow morning, you will have to look at my robots.txt file then.

33. A Web site author who wants to exclude certain crawlers from accessing all or part or his site can do so by technical means. For example, he can configure his Web server software to reject accesses (or accesses to certain parts of the site) that are labeled as coming from the targeted crawler(s)[3].

34. The requests in a robots.txt file apply, by definition, only to the initial gathering or archiving of Web sites and pages by a crawler. The requests do not apply to the subsequent storage, use, or redistribution of the archived sites and pages.

35. The requests in a robots.txt file apply, by definition, only to crawlers. They do not apply to people. Web site authors who want to exclude people from their sites use other mechanisms, such as password-protecting the site.

## III. Internet Archive's Technology

36. Internet Archive ("IA") collects an archive of the Web, containing past versions of many Web sites. To collect Web pages to store in its archive, IA uses crawlers. IA's crawlers visit Web pages from time to time, recording their contents

[3] Requests for Web pages are labeled with a "User-Agent" header saying which computer program is making the request. A Web server can reject requests if it sees the User-Agent header characteristic of a targeted crawler.

in IA's archive. IA's crawler is programmed to look for the robots.txt file on each site it visits, and to comply with any requests found therein.

37. IA also provides an online service called the Wayback Machine[4] that lets users view archived versions of sites. A user can enter the URL (Web address) of a page or site into the Wayback Machine, and the Wayback Machine will then consult IA's archive to help the user view past versions of the requested page or site.

38. IA has voluntarily chosen to have the Wayback Machine withhold archived copies of a site from users, if that site has a robots.txt file that asks the IA crawler not to visit that site. In doing this, IA goes beyond any request made by the site publisher in the robots.txt file – nothing in the robots.txt standard requires this from IA, nor does the standard indicate that doing so would be customary, polite, or otherwise desirable. A person familiar only with the robots.txt standard would have no particular reason to think that IA would do this, or that there would be anything unusual or improper about IA not doing it.

39. To implement this behavior, the Wayback Machine would have to consult a site's robots.txt file every time a user tried to look at an archived version of that site. IA apparently intended the Wayback Machine to use caching, keeping copies of already-accessed robots.txt files for twenty-four hours. If implemented correctly, this would have caused the Wayback Machine to access sites' robots.txt files less frequently.

40. During the relevant period in July 2003, the Wayback Machine's robots.txt caching mechanism was not working as intended. This is evident from the Web server logs of Healthcare Advocates ("HCA"), which show more accesses by

the Wayback Machine to HCA's robots.txt file than would have occurred had the caching mechanism been working as expected. Sometimes, when one would have expected the Wayback Machine to use a cached version of HCA's robots.txt file, the Wayback Machine instead retrieved the robots.txt file across the Internet from HCA's server.

41. The number of extra accesses is only a few hundred, a very small number by Web standards. In any case, if HCA's robots.txt file were not changing over time, failure by the Wayback Machine to cache HCA's robots.txt file should not have affected the Wayback Machine's behavior (other than causing it to retrieve the robots.txt file more often). If HCA's robots.txt file were in place, properly constructed and unchanging, and if HCA's Web server were working correctly, then the Wayback Machine should have seen that same robots.txt file on every access. In other words, the Wayback Machine caching flaw described by Mr. Lenkey cannot by itself account for the Wayback Machine's delivery to Harding Earley of the archived files at issue here.

## IV. Harding Earley's Accesses to Internet Archive

42. At issue in this litigation are a small number of accesses made by Harding Earley's employees to the Internet Archive web site.

43. These were ordinary web page accesses, made via ordinary browsers. Harding Earley's employees did not use extraordinary, unusual, sophisticated, or devious methods to access these pages. They did not engage in "hacking". Mr. Lenkey does not point to any evidence that automated tools (beyond ordinary Web browsers) were used, nor am I aware of any such evidence.

[4] Though dubbed a "Machine", this service actually uses a group of computers.

44. Even according to Mr. Lenkey, all that Harding Earley's employees did was to make ordinary accesses to publicly available Web pages. In other words, they simply asked their browser to fetch and display a web page –what every user does when browsing the web. In response to these requests, the Wayback Machine delivered the requested pages.

45. Based on descriptions available on IA's web site, one might have expected the Wayback Machine to refrain from delivering the requested pages. Nevertheless, the Wayback Machine did deliver the pages.

46. I agree with Mr. Lenkey that the reasons the Wayback Machine provided the pages cannot definitively be determined now (Lenkey Report at p. 6: "for reasons unknown").

47. We can, however, identify several likely scenarios that could have led to the observed results. First, HCA might have tinkered with its robots.txt file, so that it either did not deny access to the IA crawler or was improperly formatted in a way that confused the Wayback Machine servers. Second, HCA's web server might have been misbehaving, failing to produce the correct robots.txt file in response to the Wayback Machine's requests for it. Third, the Wayback Machine might have had trouble reading or interpreting the contents of the robots.txt file.

48. These three examples do not exhaust the possibilities, but they do give an idea of how errors by HCA or by IA could have led to the observed behavior. Regardless of what happened, Harding Earley's role was simple – to request files by the usual means, and to receive them.

49. Mr. Lenkey claims that Harding Earley's accesses to the Wayback Machine were unauthorized and that the unauthorized accesses were intentional. I disagree with his characterization of these accesses.

50. His characterization appears to rely on an assumption that Harding Earley's employees knew, or could reasonably be expected to have known, that the Wayback Machine was supposed to be preventing them from accessing the files that it was allowing them to access. This assumption seems to rest on two assertions, regarding Harding Earley seeing HCA's robots.txt file, and Harding Earley seeing an "query exclusion page" supplied by the Wayback Machine. According to Mr. Lenkey, a person seeing these two files should have known that future accesses to the requested files would be unauthorized.

51. Regarding the robots.txt file, HCA's server logs indicate that somebody at Harding Earley saw some version of HCA's robots.txt file during the relevant period, but the record does not tell us what exactly was in the version of the robots.txt file that Harding Earley saw. In any case, as explained above in paragraph 35, the robots.txt file at most requested that *IA's robot* not visit HCA's site. HCA does not assert that the robots.txt file requested that *Harding Earley* refrain from accessing HCA's site. In any case, the robots.txt file, whatever it contained, could not have denied permission to any *person*.

52. According to Mr. Lenkey, the relevant portion of HCA's robots.txt file consisted of these two lines of cryptic text:

```
User-agent: ia_archiver
Disallow: /
```